# PixArt-$\alpha$: Fast Training of Diffusion Transformer for Photorealistic Text-to-Image Synthesis

Year: 2023 | Citations: 632 | Authors: Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie

---

## Abstract

The most advanced text-to-image (T2I) models require significant training costs (e.g., millions of GPU hours), seriously hindering the fundamental innovation for the AIGC community while increasing CO2 emissions. This paper introduces PIXART-$\alpha$, a Transformer-based T2I diffusion model whose image generation quality is competitive with state-of-the-art image generators (e.g., Imagen, SDXL, and even Midjourney), reaching near-commercial application standards. Additionally, it supports high-resolution image synthesis up to 1024px resolution with low training cost, as shown in Figure 1 and 2. To achieve this goal, three core designs are proposed: (1) Training strategy decomposition: We devise three distinct training steps that separately optimize pixel dependency, text-image alignment, and image aesthetic quality; (2) Efficient T2I Transformer: We incorporate cross-attention modules into Diffusion Transformer (DiT) to inject text conditions and streamline the computation-intensive class-condition branch; (3) High-informative data: We emphasize the significance of concept density in text-image pairs and leverage a large Vision-Language model to auto-label dense pseudo-captions to assist text-image alignment learning. As a result, PIXART-$\alpha$'s training speed markedly surpasses existing large-scale T2I models, e.g., PIXART-$\alpha$ only takes 10.8% of Stable Diffusion v1.5's training time (675 vs. 6,250 A100 GPU days), saving nearly \$300,000 (\$26,000 vs. \$320,000) and reducing 90% CO2 emissions. Moreover, compared with a larger SOTA model, RAPHAEL, our training cost is merely 1%. Extensive experiments demonstrate that PIXART-$\alpha$ excels in image quality, artistry, and semantic control. We hope PIXART-$\alpha$ will provide new insights to the AIGC community and startups to accelerate building their own high-quality yet low-cost generative models from scratch.